

# Standard deviations and correlations of GC levels in DNA sequences

Oliver Clay\*

*Laboratory of Molecular Evolution, Stazione Zoologica “Anton Dohrn”, Villa Comunale, 80121 Naples, Italy*

Received 12 May 2001; received in revised form 23 June 2001; accepted 10 August 2001

Received by C.W. Schmid

## Abstract

In a DNA sequence that exhibits long-range correlations, standard deviations among the GC levels of its segments can be up to an order of magnitude higher than in a sequence consisting of independent, identically distributed nucleotides. Conversely, plots of inter-segment standard deviations vs. segment length reveal quantitative information about the correlations present in a sequence. We present and discuss formulae that relate long-range (power-law) correlations between the nucleotides of a sequence to the expected standard deviations of the GC levels of its segments, and to the correlations between them. © 2001 Elsevier Science B.V. All rights reserved.

*Keywords:* Analytical ultracentrifugation; Base composition; DNA; Genomes; Isochores; Long-range correlations

## 1. Introduction

Long-range correlations between nucleotides, or between longer regions of a DNA sequence, are characterized by a slow decrease with respect to the intervening distance between the regions. The prototype of such correlations is the power-law form of the correlation function, which can be used to characterize, at least approximately or asymptotically, the class of long-range correlated DNA sequences. The existence of power-law correlations allows, in turn, formulae to be derived for the expected standard deviation among fixed-length regions within the sequences. We here present and discuss formulae that relate power-law correlations between nucleotides to the standard deviations and correlations between longer regions to which they lead. The main results remain good approximations even when the correlations do not obey a perfect power-law for all distances, as in the case of real DNA sequences (see the empirical results in Clay et al., 2001; Clay and Bernardi, 2001). For derivations when the power-law requirement is relaxed, as well as for a more general and extended treatment of long-range correlations, we refer to Beran (1994) and to the references cited therein.

In what follows, we will assume that the DNA sequences

of interest can be described by well-defined correlation functions. We therefore first review conditions for a correlation function to satisfactorily model a DNA sequence, and interpret them in the context of a genome organized into isochores. In so doing, we will sometimes regard sequences of GC/AT base pairs as if they were sequences of binary digits (1/0), or, equivalently, as if they were runs, or time series, of consecutive tosses of a coin (head/tail; Bernoulli trial).

## 2. Correlated Bernoulli trials and DNA sequences

A trial with two possible outcomes, 1 (head of a coin, GC base pair) or 0 (tail of a coin, AT base pair), is called a Bernoulli trial; a random variable that takes these two values is called a Bernoulli random variable. Whereas for the case of uncorrelated (independent) Bernoulli trials a well-developed theory has existed since 1812 at the latest (Jakob Bernoulli, De Moivre, Laplace), and is discussed in most elementary textbooks on probability or statistics, the case of correlated (dependent) Bernoulli trials, which is the relevant case for statistical studies of nucleotide sequences, has received surprisingly little attention (Viveros et al., 1994). This is apparently the situation even for power-law correlations, simple as they are to define. In the limit of many trials (runs of  $l$  trials, long DNA segments of length  $l$ ), the frequency distributions to which they correspond (number or percentage of 1s, or percentage of GC base pairs) usually become approximately Gaussian, with a standard deviation that will be given below. In general, however, no approx-

Abbreviations: bp, base pairs; kb, kilobase pairs; Mb, megabase pairs; GC, molar fraction of guanine and cytosine in DNA;  $\sigma$ , standard deviation;  $\propto$ , proportional to;  $\approx$ , approximately equal to;  $O(\dots)$ , of the order of;  $\langle \dots \rangle$ , mean of;  $|\dots|$ , absolute value of

\* Fax: +39-081-764-1355.

E-mail address: clay@infobiogen.fr (O. Clay).

imate formulae appear to be available for the frequency distributions (GC distributions) or their third moments (asymmetries), only for their first two moments (mean, standard deviation).

### 3. Correlation functions

A correlation function  $c(d)$  specifies the (expected) correlation between two base pairs as a function of the distance  $d$  between them. For such a correlation function to be a meaningful and useful description of a collection of DNA sequence(s), at least three conditions should be met. Statistical properties of GC levels should not vary along the sequence(s); an ergodic condition, closely related to the previous condition, should be satisfied, and the correlation function should be self-consistent.

The first of these three conditions is usually called (spatial) stationarity along the sequence. In order to avoid the possibly confusing temporal connotations of this term (which does not have an established synonym), we will instead refer to positional invariance of the GC distribution along a sequence. Indeed, much of the theoretical work on correlations was developed in the temporal context of time series, but its definitions and results can be easily re-interpreted in a spatial context: in our case, we will re-interpret time as the position along a DNA sequence or chromosome. A time unit will correspond to the distance between two successive nucleotides or, later, between two successive segments of a given length.

The positional invariance condition that we will require is that a sequence's GC level and its variation be adequately described by a single mean value, and by an inter-nucleotide correlation function that is valid throughout the sequence. As we shall see below, in this case one can calculate an expected standard deviation among fixed-length segments that will also remain valid throughout the sequence. A stricter condition would require that the entire expected GC distribution of the sequence's segments at any given length should remain the same throughout the sequence, i.e. that not just the expected distributions' means and standard deviations, but also their shapes, should remain identical. (In the technical literature, our basic condition is sometimes called 'wide-sense' stationarity, and the stricter condition, 'strict-sense' stationarity (Shiryayev, 1984, p. 387).)

Chromosomes are mosaics of isochores, and thus do not, in general, fulfil the above condition. In contrast, individual isochores do usually fulfil the condition, and in some cases they even fulfil the stricter property of invariance of GC distributions along the sequence, as is illustrated for an isochore of chromosome 21 in Clay et al. (2001, Fig. 2).

An imaginary sequence consisting of concatenated isochores from a single isochore family could again be regarded as fulfilling the above condition, since the isochores will have very similar statistical properties (cf. Clay et al., 2001). Another way of describing this statistical

similarity within families would be to say that each family satisfies an 'ergodicity' property. In statistical physics and time series analysis, one distinguishes between the 'time average' of a quantity or observable of interest and its 'ensemble average', taken over an ensemble or set of instances or examples of a system that can be assumed to have arisen in a similar way: an 'ergodic condition' is fulfilled if the time average of a quantity of interest is equal to its ensemble average. In our re-interpretation of this scenario, time corresponds to a position along a DNA sequence, time averages correspond to position averages along an isochore or contig, and ensembles correspond to isochore families, since we assume that similar selection pressures gave rise to the isochores of the same family, even if their locations are dispersed throughout the genome. We require, then, that the averages of statistical parameters involving GC levels, taken over a single isochore (corresponding to 'time averages' in the temporal context) should be essentially the same as those taken over the entire isochore family to which it belongs (corresponding to 'ensemble averages'). The ergodic condition therefore extends the above condition of positional invariance, which we had formulated for individual sequences, to cross-sequence comparisons. Thus, averages (i.e. expected values) of mean, correlation function and standard deviation, as estimated from a collection of fixed-length segments, should be essentially the same whether the segments are taken from a single isochore or from different isochores of the same isochore family. We may then speak of a correlation function being valid throughout an isochore family, and analyze collectively the pooled sequences from different isochores of the same family.

Finally, the self-consistency (non-degeneracy) property ensures that the correlation function is well-defined. For example, a strong positive correlation between adjacent base pairs that is valid throughout a sequence,  $c(1) \approx 1$ , implies, by transitivity, a moderately strong positive correlation also between next-to-nearest neighbors, so that  $c(2)$  cannot be less than  $2c(1)^2 - 1$ . Any postulated correlation function  $c(d)$  must, therefore, first of all be compatible with such constraints, or it will lead to internal inconsistencies: a putative correlation function that is abstractly defined by  $c(1) = 0.9$  for adjacent nucleotides, yet by  $c(d) = 0$  for all larger inter-nucleotide distances  $d > 1$ , is self-contradictory. The general condition is that the theoretical correlation matrix, or Laurent matrix,  $(c(i-j))_{i,j}$ , where  $i, j = 1, 2, \dots, l$ , must remain non-negative-definite for all subsequences or fragments of length  $l$  (e.g. Kendall and Stuart, 1976, pp. 424–428; Shiryayev, 1984, p. 233). For this condition to be valid, no minor along the main diagonal of the Laurent matrix may have a negative determinant. Interestingly, already a very simple power-law correlation function, given by  $c(d) = d^{-1/4}$  for  $d > 0$  and  $c(0) = 1$ , violates this condition, since any  $4 \times 4$  minor along the main diagonal of its Laurent matrix is negative. It can, however, be made self-consistent by a slight adjustment, namely if  $d$  is replaced by  $d + 1$ .

The fulfillment of the three conditions listed here encourages the concept of a single correlation function, approximately valid along all sequences from a given isochore family, to be employed, and its consequences to be explored.

#### 4. Standard deviation of the GC distribution for an ideal power-law correlation

Self-consistency (non-degeneracy) is fulfilled by a class of power-law correlation functions that differ only marginally from the type mentioned above: those having the general form  $c(d) = a(d+1)^{-\gamma}$ , where  $d > 0$ , and both the factor  $a$  and the exponent  $\gamma$  lie between 0 and 1. For simplicity, we will make the idealizing assumption that this equation holds perfectly for all  $d$ . It is then straightforward to show that the standard deviation  $\sigma$  of the fragment distribution is, for sequences governed by such a correlation function, in turn a power-law function of fragment length, with power  $-\beta = -\gamma/2$ .

In practice this result, and the main results presented below, will remain valid even when a function of the above class approximates the true correlation only roughly; rigorous results extending some of the relations to less ideal situations can be found in Beran (1994) (where a parameter  $H \equiv 1 - \gamma/2$  is used). Indeed, Beran (1994) gives an asymptotic definition of long-range correlations that requires a power-law behavior of the correlations only when the lag  $d$  is large (tends to infinity).

Let  $x_p \equiv (1/l)(u_p + u_{p-1} + \dots + u_{p-l+1})$  be the GC content, expressed as a fraction of 1, of a fragment of DNA of length  $l$  that ends at nucleotide (position)  $p$ , where  $u_j$  is 1 if there is a G or C at position  $j$ , or else 0. For a sequence that can be described by a self-consistent correlation function

$$c(d) \equiv \langle (u_i - \mu)(u_{i+d} - \mu) \rangle / (\mu(1 - \mu)) \\ = \langle (u_i u_{i+d}) - \mu^2 \rangle / (\mu(1 - \mu))$$

we can directly calculate the variance among the fragments, obtaining

$$\sigma^2 = \langle (x_p - \mu)^2 \rangle = \frac{\mu(1 - \mu)}{l} \left( 1 + 2 \sum_{d=1}^{l-1} \left(1 - \frac{d}{l}\right) c(d) \right) \quad (1)$$

where  $\mu = \langle u_p \rangle \approx \langle x_p \rangle$  is the mean GC of the full sequence. This relation holds for any form of the correlation  $c(d)$ , including the familiar correlation-free (binomial) case in which  $c(d) = 0$  for all  $d > 0$ ,

$$\sigma_{\text{uncorrelated}}^2 = \frac{\mu(1 - \mu)}{l}. \quad (2)$$

For the power-law correlation  $c(d) = a(d+1)^{-\gamma}$ , with  $\gamma$  around 0.3–0.5, we can approximate the sum from 1 to  $l-1$  in Eq. (1) by an integral from 0 to  $l$ . This yields, when  $l$  is large and the contribution from higher order terms is small,

the desired expression for the variance among the fragments of length  $l$ ,

$$\sigma^2 \approx a \frac{\mu(1 - \mu)}{(1 - \beta)(1 - 2\beta)} l^{-2\beta} \quad (3)$$

where  $\beta = \gamma/2$ . (Further details are given in Beran (1994); similar calculations can also be found in Taqqu et al. (1995).)

Thus, the standard deviation  $\sigma$  of the GC distribution is, to a good approximation, a power-law function of the fragment length  $l$ , the exponent  $\beta$  being half that of the inter-nucleotide correlation function  $c(d)$ . For  $\beta = 0.15$ , 50% GC and  $l = 250$  bp, the relative error of this power-law approximation is less than 5%; for  $l = 1$  kb it is less than 2%. A double-logarithmic plot of  $\sigma$  vs.  $l$  should, therefore, essentially follow a straight line of the form  $\log \sigma = -\beta \log l + b$ , as is indeed observed in human isochores and isochore families (Clay et al., 2001, Fig. 3).

When fragment sizes increase, the form of the GC distribution usually approaches a Gaussian, with a standard deviation given by Eq. (3). In DNA, a power-law relation of the form in Eq. (3) is reached already for fragment lengths  $l$  of about 100 bp, while the Gaussian form is not reached until fragments are longer ( $l > 1$  kb). For short-fragment GC distributions of an isochore family, or of a relatively homogeneous genome such as that of a bacterium, the asymmetry tends to be positive when the mean GC level is less than about 42–44% GC, and negative when it is higher. (Early examples of this tendency, which we have since confirmed by CsCl and sequence analyses of bacteria and mammalian isochore families, were found by Yamagishi (1970, 1971, 1974).)

In large-scale analyses of DNA we mainly consider fragment lengths above 100 bp. For simplicity we then often refer, neglecting rigor, to the nearly self-consistent approximation  $ad^{-\gamma}$ , rather than to the more cumbersome exact function  $a(d+1)^{-\gamma}$ . We will similarly neglect rigor in treating  $\beta$  and  $\gamma/2$  as identical, even though, as we have seen, this is an approximation, and only valid for the fragment lengths and exponent values in the ranges that interest us.

In this context, a caveat should, however, be mentioned: the approach of the standard deviation to a power-law behavior can be perturbed when correlations  $c(d)$  deviate markedly from a power-law behavior, especially where distances  $d$  are short. Since a correlogram for DNA cannot be expected to be a perfect power function valid for all  $d$ , the accuracy in estimating  $\gamma$  from the relation  $\beta \approx \gamma/2$  will not always be as high as suggested by the numerical examples given above.

#### 5. Correlation between adjacent non-overlapping segments for an ideal power-law correlation

Calculations, analogous to those above, lead to an exact formula for the correlation  $C(1)$  between adjacent segments

of length  $l$ , in terms of the power-law correlation  $c(d)$  between individual sites (nucleotides):

$$C(1) = \frac{1}{\sigma^2} \frac{\mu(1-\mu)}{l^2} \left( lc(l) + \sum_{d=1}^{l-1} d(c(d) + c(2l-d)) \right) \quad (4)$$

Integration again yields the corresponding power-law approximation when the inter-nucleotide correlation function is  $c(d) = a(d+1)^{-2\beta}$ :

$$C(1) \approx 2^{1-2\beta} - 1 \quad (5)$$

Here and in the following, we shall use lower-case letters,  $c(d)$ , to denote inter-nucleotide correlations as a function of the distance between their positions, and upper-case letters,  $C(n)$ , to denote inter-segment correlations as a function of the number of segments between their positions.

It is interesting that the approximation in Eq. (5) of the adjacent-segment correlation, which becomes accurate when  $l$  is moderately large (above about 500 bp), is not dependent on  $l$ ,  $a$  or  $\mu$ . Its lack of dependence on  $\mu$  is only a formal one, since  $\beta$  decreases with increasing mean GC level  $\mu$ , up to at least 50–55% GC. On the other hand, its lack of dependence on  $a$  and  $l$  is noteworthy.

The disappearance of the fragment length  $l$  from Eq. (5) is a remarkable result, because it implies complete scale-invariance ('scaling'): the expected correlation between adjacent segments should in principle be the same whether one compares adjacent 1 kb segments or adjacent 10 kb segments. Indeed, if the power-law correlation function  $c(d)$  were perfectly valid for all distances  $d$  (an ideal, unrealistic situation), there would be no scale beyond which one could neglect its large-scale effects. This property gives an intuitively obvious meaning to descriptions of power-law correlations as 'long-range' or 'large-scale'.

As a numerical example, we consider an exponent  $\beta = 0.15$ , which approximates the behavior of the standard deviation in GC-rich human isochores (Clay et al., 2001), and leads to an adjacent-segment correlation  $C(1)$  of 0.625. This value is, therefore, the approximate correlation coefficient  $R$  that we expect between the GC levels of adjacent segments from GC-rich isochores, e.g. as visualized in a scatterplot showing GC levels of pairs of adjacent segments ('time-delay plot', 'phase plot'; cf. Clay and Bernardi, 2001).

When the segments from all isochores of a genome are included in such plots (ignoring for now that the conditions of Section 3 may not be fulfilled), one can obtain higher, although again largely scale-invariant, correlation coefficients  $R$ . For example, in Jabbari and Bernardi (2000, Fig. 4), human fragment lengths of  $l = 50$  kb and  $l = 100$  kb yield correlation coefficients of 0.83–0.88. In this case, not only the intra-isochore correlations  $C(1)$ , but also the mosaic structure of the genome, contributes to the final value of  $R$ , since the clouds of points representing isochore families are spread out along the main diagonal of slope 1. Similar scale-invariant correlation coefficients would, interestingly, be expected also for a sequence that is character-

ized by  $\beta \approx 0.06$ , i.e. by the local slope of the  $\sigma$  vs.  $l$  plot actually observed for the total human DNA where  $l \approx 50$ –100 kb (Clay et al., 2001, Fig. 3). Indeed, DNA exhibiting obvious mosaicism at the scale of a few isochores (i.e. having different mean GC levels, and thus violating the conditions for a single correlation function to be meaningful) can, at much larger scales, be viewed as having a single mean and a single long-range correlation function. In some cases, the scales at which this view becomes appropriate would, however, be longer than the chromosome.

Returning to the isochore scale, we should mention that it is not always easy to verify scale-invariance of the correlation in Eq. (5) within a single isochore. Isochores often do not have perfectly invariant statistical properties throughout their length, are not governed by perfect power-law correlations, and are not long enough to allow comparison of adjacent-segment correlations for fragment lengths spanning much more than one order of magnitude: although GC-poor isochores are often longer than GC-rich isochores, their fluctuations and expected correlations are lower, and sample sizes quickly become insignificant in both cases. Nevertheless, in the long,  $\approx 7$  Mb, GC-poor isochore of chromosome 21, the empirical correlation remains almost constant, and significant, for fragments ranging from 1 to 50 kb. Its value, 0.16–0.19, remains close to the value expected from Eq. (5), when one uses the slope of the  $\sigma$  vs.  $l$  plot for this isochore (Clay et al., 2001, Fig. 4) to estimate  $\beta$ .

## 6. Correlation between distant non-overlapping segments for an ideal power-law correlation

The above calculations can be easily generalized to obtain the correlation  $C(n)$  between non-overlapping segments of the same length  $l$  whose positions are separated by a distance of  $n$  segments. The exact formula is identical to Eq. (4), except that the arguments of the correlation function are generalized:  $l \rightarrow nl$ ,  $d \rightarrow (n-1)l + d$ ,  $2l - d \rightarrow (n+1)l - d$ . The corresponding approximation for power-law correlations is then

$$C(n) = \frac{1}{2} \left( (n+1)^{2-2\beta} - 2n^{2-2\beta} + (n-1)^{2-2\beta} \right) \quad (6)$$

(This equation has been proposed as a definition of 'exact self-similarity' (Paxson and Floyd, 1995, and references therein; cf. also Beran, 1994, p. 52).)

In the distant-segment limit ( $n$  large) we obtain

$$C(n) = \frac{(1-\beta)(1-2\beta)}{2} n^{-2\beta} \left( 1 + O(n^{-2}) \right) \quad (7)$$

For very distant segments, where the term of order  $n^{-2}$  can be neglected, the inter-segment correlation as a function of the number of segments is again a power-law function, with the same exponent as the original inter-nucleotide correlation function. In other words, 'aggregation' of the sequence into segments does not change the exponent. (Different

‘scaling’ properties of DNA and of polymers in general can be found in de Gennes (1979)).

## 7. Exponents of correlation, standard deviation and Fourier spectra

By using the Wiener–Khinchin theorem, one can show that a binary symbol sequence having a Fourier power spectrum of the form  $S(f) = |A_f|^2 \propto f^{-\alpha}$  with  $\alpha$  between 0 and 1, is governed by an inter-symbol correlation function of the form  $c(d) \propto d^{\alpha-1}$ , and vice versa (Li, 1997; Talenti, 1995; Beran, 1994). In other words, Fourier amplitudes  $|A_f| \propto f^{-\alpha/2}$  that are power-law functions of the frequency  $f$  correspond to correlation functions  $c(d)$  that are power-law functions of the distance  $d$ . Thus, we can use any one of three different power-laws to describe the same phenomenon, observed in DNA:  $\sigma(l) \propto l^{-\beta}$ ,  $c(d) \propto d^{-\gamma}$ , and  $S(f) \propto f^{-\alpha}$ , where  $\alpha \approx 1 - \gamma \approx 1 - 2\beta$ . For example,  $\beta = 0.15$  corresponds to  $\gamma \approx 0.3$  and  $\alpha \approx 0.7$ . As the exponent  $\alpha$  approaches 1 (perfect 1/f noise), the exponents  $\gamma$  and  $\beta$  should approach 0, so that log–log plots of  $\sigma$  vs.  $l$  should become nearly horizontal.

## 8. Remarks on window overlap

A moving window plot of a DNA sequence, obtained by plotting the GC level of each window position over the midpoint of the window, is a GC level plot of possible fragments. A question that may appear relevant in the context of other papers presented in this issue (Clay et al., 2001; Clay and Bernardi, 2001) is which degree of overlap (step size between successive positions of the window) should be used in order to best simulate the fragment collections represented by experimental CsCl profiles, i.e. by GC distributions obtained via density gradient ultracentrifugation.

A quick back-of-the-envelope calculation shows that a sample of a species’ DNA that consists of random DNA fragments of the same length should be roughly equal to the non-random collection of all possible fragments (subsequences) of that length, since the latter fragment set is what a window of that length yields, when it has moved through the complete genomic sequence with a step size of 1 bp. Indeed, a typical preparation of mammalian DNA used for obtaining CsCl profiles by analytical ultracentrifugation contains about  $5 \times 10^4$ – $5 \times 10^5$  cells (Macaya et al., 1976), while the highest molecular weight preparations that can be reliably studied by CsCl gradient ultracentrifugation have lengths of about  $5 \times 10^4$ – $5 \times 10^5$  bp (cf. Macaya et al., 1976). Thus, if in each cell of the preparation the genome were fragmented or broken at different positions, each possible fragment of a given size would be represented approximately once in the sample. The choice of ‘overlapping’ windows (which, as we see here, does not correspond to an overlap in any of the samples’ cells) is not only realis-

tic, but it also has the advantage that it yields smoother histograms, encoding more information on the sequences (see, for example, Clay et al., 2001, Fig. 2).

The distinction between overlapping and non-overlapping fragments is crucial when calculating a correlation between the GC levels of successive fragments of a correlated sequence, e.g. when deriving Eq. (4). On the other hand, when calculating the standard deviation of the fragments’ GC levels, this distinction is largely academic. If we consider a large collection of identical-length fragments covering a long DNA sequence (long compared to its typical GC fluctuations, to ensure a representative sample, and long compared to the fragments’ lengths, to ensure a large non-redundant sample), then their standard deviation will be independent of their overlap. In the uncorrelated case it will always be given by Eq. (2), and in the correlated case by Eq. (1). If this seems counter-intuitive at first, it can be easily understood by imagining the sequence first partitioned into a cover of non-overlapping segments, then into another non-overlapping cover in which all segments are displaced with respect to the first: the two covers will obviously have standard deviations that are similar to each other, and therefore also to that of the overlapping cover obtained when their segments are pooled.

## Acknowledgements

I am indebted to Giorgio Bernardi, Gabriel Macaya, Wentian Li, José Oliver, Pedro Bernaola-Galván, Pedro Carpena and Carl Schmid for helpful discussions, comments and suggestions.

## References

- Beran, J., 1994. Statistics for Long-Memory Processes. Chapman and Hall/CRC, Boca Raton, FL.
- Clay, O., Bernardi, G., 2001. Compositional heterogeneity within and among isochores in mammalian genomes. II. Some general comments. *Gene*, 276, 25–31.
- Clay, O., Carels, N., Douady, C., Macaya, G., Bernardi, G., 2001. Compositional heterogeneity within and among isochores in mammalian genomes. I. CsCl and sequence analyses. *Gene*, 276, 15–24.
- de Gennes, P.-G., 1979. Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca, NY.
- Jabbari, K., Bernardi, G., 2000. The distribution of genes in the *Drosophila* genome. *Gene* 247, 287–292.
- Kendall, M., Stuart, A., 1976. The Advanced Theory of Statistics, 3rd Edition. Charles Griffin, London.
- Li, W., 1997. The study of correlation structures of DNA sequences: a critical review. *Comput. Chem.* 21, 257–272.
- Macaya, G., Thiery, J.P., Bernardi, G., 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* 108, 237–254.
- Paxson, V., Floyd, S., 1995. Wide-area traffic: the failure of Poisson modeling. *IEEE/ACM Trans. Networking* 3, 226–244.
- Shiryayev, A., 1984. Probability. Springer-Verlag, New York.
- Talenti, G., 1995. Trasformazione integrale. *Enciclopedia delle Scienze Fisiche*, Istituto della Enciclopedia Italiana, Rome, pp. 297–305.

- Taqqu, M., Teverovsky, V., Willinger, W., 1995. Estimators for long-range dependence: an empirical study. *Fractals* 3, 785–788.
- Viveros, R., Balasubramanian, K., Balakrishnan, N., 1994. Binomial and negative binomial analogues under correlated Bernoulli trials. *Am. Stat.* 48, 243–247.
- Yamagishi, H., 1970. Nucleotide distribution in the DNA of *Escherichia coli*. *J. Mol. Biol.* 49, 603–608.
- Yamagishi, H., 1971. Heterogeneity in nucleotide composition of *Bacillus subtilis*. *J. Mol. Biol.* 57, 369–371.
- Yamagishi, H., 1974. Nucleotide distribution in bacterial DNAs differing in G + C content. *J. Mol. Evol.* 3, 239–242.